

Lec: 6

[2] Classification : discrete valued output.

① Binary Classification.

Has two different classes.

ex: Email : Spam / Not Spam

Tumor : Malignant / Benign

y : output $\in \{0, 1\}$ binary classification problem

we can say $y \in \{1, 2\}$ is binary classification problem

Here, binary means 2 different classes.

0: "Negative Class" (eg benign tumor, spam email)

1: "Positive Class" (eg malignant tumor, not spam email)

② Multiclass Classification.

Has more than 2 different classes.

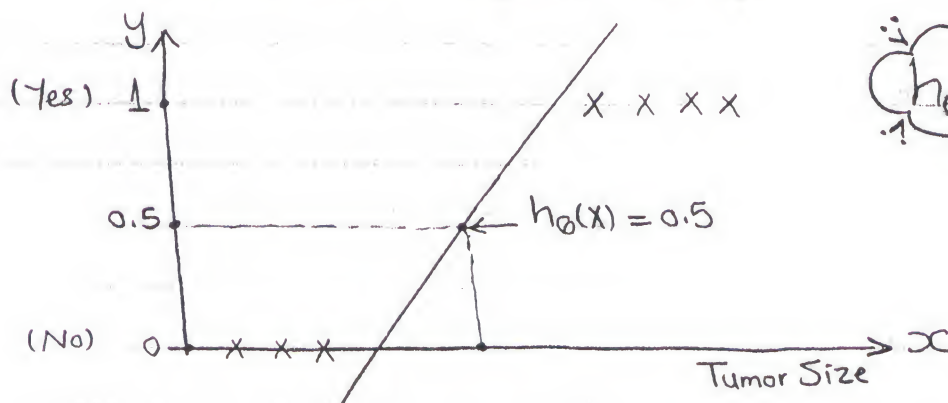
\hat{y} is the predicted variable.

$\hat{y} \in \{0, 1, 2, 3\}$ is a multiclass classification problem.

Using Linear Regression to get good Fitting to the data Set

we explain by example

ex: Tumor : Malignant / Benign.



$$h_0(x) = \theta^T x$$

In this example with this data set we can get good Fitting to the data

1. We determine the threshold classifier output

→ Here, Threshold Classifier output $h_0(x)$ at 0.5

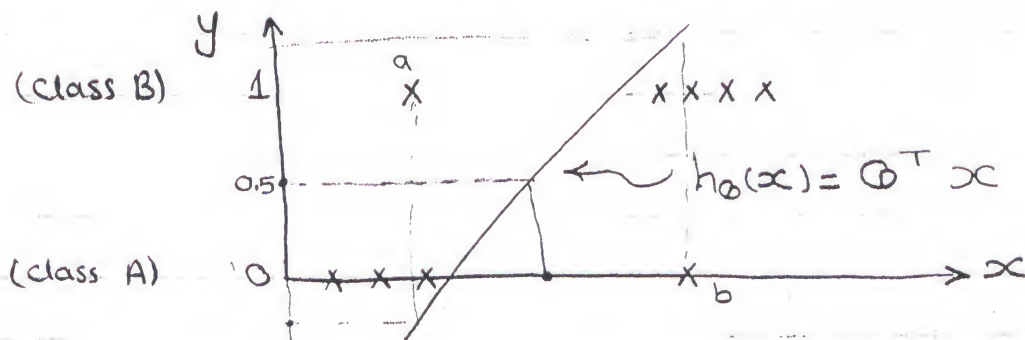
- Note: 1. Threshold Classifier output is Very Important to Know the Class in which every point (data set) belong
- 2. This point can take any Value (not necessarily mean Value)
القيمة المتوسطة
- We Choose the value that give us the best Fitting.

→ From the threshold Classifier output $h_0(x)$ at 0.5
We divide the data sets into two classes depending on the value of $h_0(x)$:

IF $h_0(x) \geq 0.5$, predict : " $y = 1$ "

IF $h_0(x) < 0.5$, predict : " $y = 0$ "

→ IF we have another data set



هذا مثال على
خط التفرقة
الذي يفصل بين
البيانات الخطية

→ For a $h_0(x) < 0.5$ this mean that it belongs to class A
but it belongs to class B in Fact.

→ For b $h_0(x) > 0.5$ this mean that it belongs to class B
but it belongs to class A in Fact.

Note : Using Linear regression, the hypothesis can output values much
Longer than one or less than zero even if all of good the
training examples have Labels y equals zero or one

• Generally, Linear Regression Can't give good Fitting to the
data set in all cases. 32.

Logistic Regression

→ We have a problem using Linear regression when we get values of $h_0(x) > 1$ and values of $h_0(x) < 0$

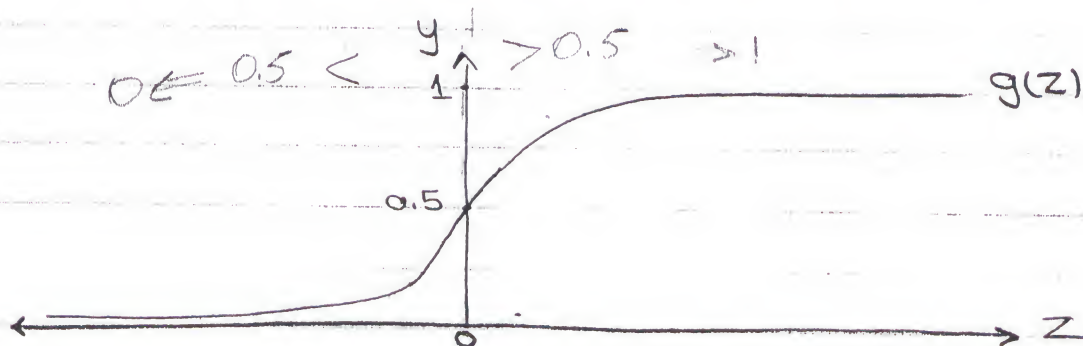
→ Want $0 \leq h_0(x) \leq 1$

$$h_0(x) = g(\theta^T x)$$

Where Linear regression model $h_0(x) = \theta^T x$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Where $g(z)$: Sigmoid Function OR Logistic Function.



$$\text{at } z \rightarrow \infty \quad g(z) = 1$$

$$\text{at } z \rightarrow -\infty \quad g(z) = 0$$

Interpretation OF Hypothesis Output

$h_0(x)$ = estimated probability that $y=1$ on input x .

$$0 \leq h_0(x) \leq 1$$

probability of the input being belongs to one of 2 classes.

$$\text{ex: IF } x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$$

$$\text{IF we get } h_0(x) = 0.7$$

This mean that 70% chance of tumor being malignant.

$h_{\theta}(x) = P(y=1 | x; \theta)$
 "probability that $y=1$, given x , parameterized by θ "

∞ $y=0$ or $y=1$

∞ $P(y=0 | x; \theta) + P(y=1 | x; \theta) = 1$

OR

$P(y=0 | x; \theta) = 1 - P(y=1 | x; \theta)$

* ————— *

• Decision boundary

$\rightarrow h_{\theta}(x) = g(\theta^T x)$, $g(z) = \frac{1}{1+e^{-z}}$

\rightarrow Where $\theta^T x$: Equation.

• we will choose the best equation that give us the best fitting

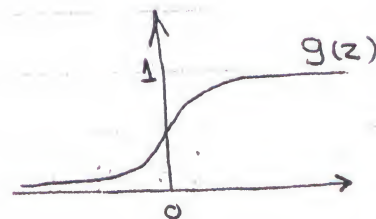
Predict " $y=1$ " IF $h_{\theta}(x) \geq 0.5$

• $g(z) \geq 0.5$ when $z \geq 0$

• $h_{\theta}(x) \geq 0.5$ when $x \geq 0$

$\therefore h_{\theta}(x) = g(\theta^T x)$

• $g(\theta^T x) = h_{\theta}(x) \geq 0.5$ whenever $\theta^T x \geq 0$



Predict " $y=0$ " IF $h_{\theta}(x) < 0.5$

• $g(z) < 0.5$ when $z < 0$

• $g(\theta^T x) = h_{\theta}(x) < 0.5$ whenever $\theta^T x < 0$

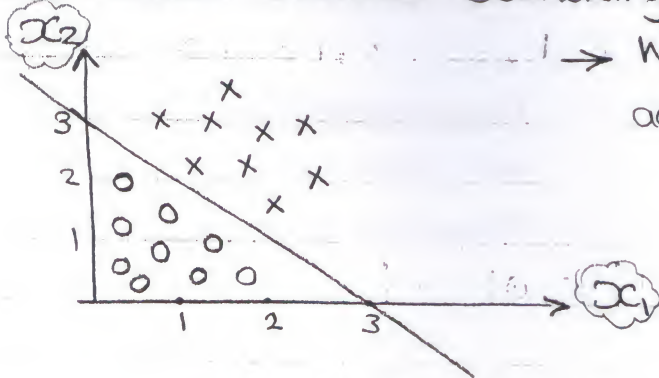
So, we can get the equation $\theta^T x$ and Apply the two conditions

$y=1 \rightarrow \theta^T x \geq 0$

$y=0 \rightarrow \theta^T x < 0$

we can replace the $\theta^T x$ by the parameters of the equation.

1 Linear Decision Boundary



→ We choose the equation of $\Phi^T x$ according to the data set.

→ According to this data set we choose the Straight Line to Fit it, so equation of $\Phi^T x$ is 1st order equation.

$$h_{\Phi}(x) = g(\Phi_0 + \Phi_1 x_1 + \Phi_2 x_2)$$

$$\therefore \Phi^T x = \Phi_0 + \Phi_1 x_1 + \Phi_2 x_2 = x_1 + x_2 = 3$$

$$\text{Assuming } \Phi_0 = -3, \Phi_1 = \Phi_2 = 1$$

→ Predict 'y=1' IF $\Phi^T x \geq 0$

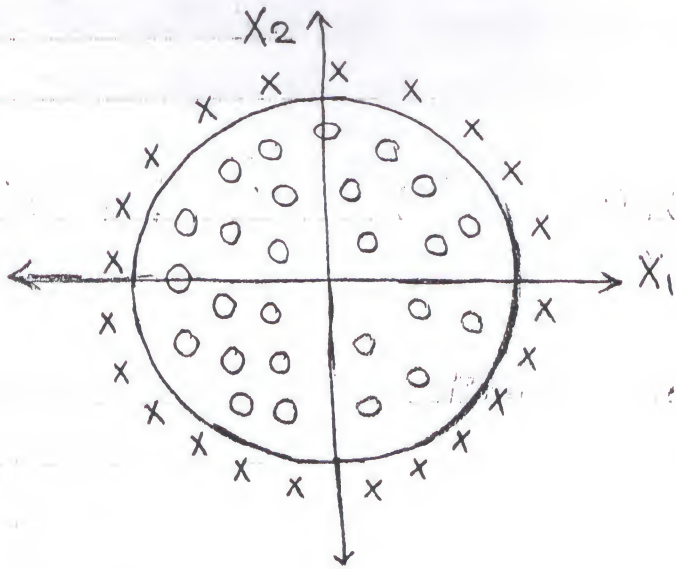
where y=1 when $g(\Phi^T x) = h_{\Phi}(x) \geq 0.5 \Rightarrow \Phi^T x \geq 0$

Predict 'y=1' IF $-3 + x_1 + x_2 \geq 0$

This mean for any example of features x_1 and x_2 that satisfy this equation $-3 + x_1 + x_2 \geq 0 \Rightarrow y=1$

$\therefore x_1 + x_2 = 3$ is the decision boundary where $h_{\Phi}(x) = 0.5$.

② Non. Linear Decision Boundary.



Here we choose circle to fit this data set, we also can use ellipso to do the same thing

Polynomial Regression: add extra higher order polynomial Features For the hypothesis.

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Assuming $\theta_0 = -1$, $\theta_1 = \theta_2 = 0$, $\theta_3 = \theta_4 = 1$ على مسان او ميل لهندسة الدائرة

$$\therefore \text{Predict "y=1" if } -1 + x_1^2 + x_2^2 \geq 0$$

* ————— *

In another data set we add extra higher order polynomial

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1^3 \dots)$$

• Logistic Regression

• Cost Function

training set : $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad m \text{ examples}$$

$$x_0 = 1$$

$$y \in \{0, 1\}$$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \rightarrow z$$

→ at Ideal Case Cost Function = Zero

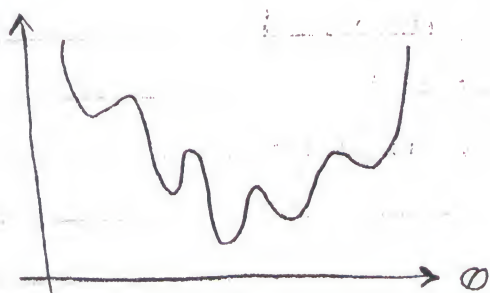
• $h_\theta(x) = y = 1$ For data set \in Class A

• $h_\theta(x) = y = 0$ For data set \in Class B

→ We choose parameters θ given this training sets which give us small value of Cost Function to get the best model fitting this data set.

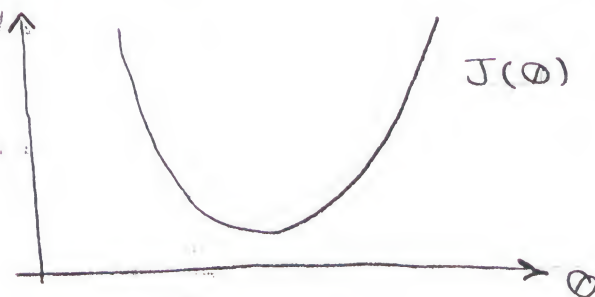
Linear regression: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Logistic regression: $\text{Cost}(h_{\theta}(x), y) = \frac{1}{2} (h_{\theta}(x) - y)^2$



Not-Convex.

• more than one local minimum.

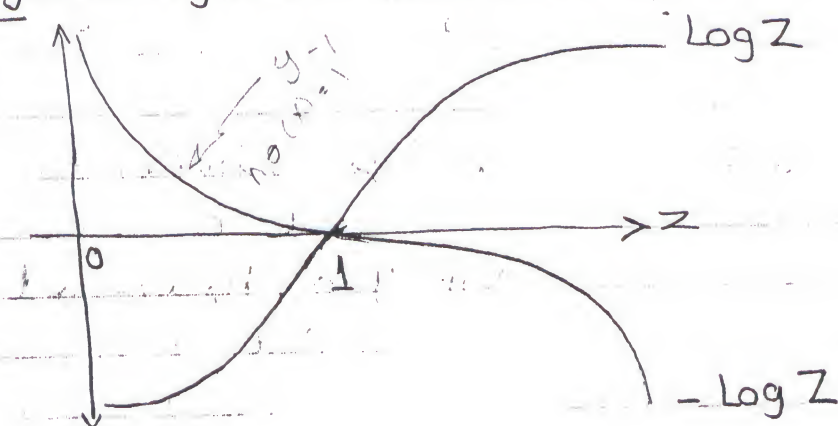


Convex

• one global minimum.

→ We choose Convex Cost Function when using Logistic regression, as it has one global minimum. (this help us to get the smallest Cost Function Value to get the best model Fitting our data set.)

Logistic regression Cost Function.



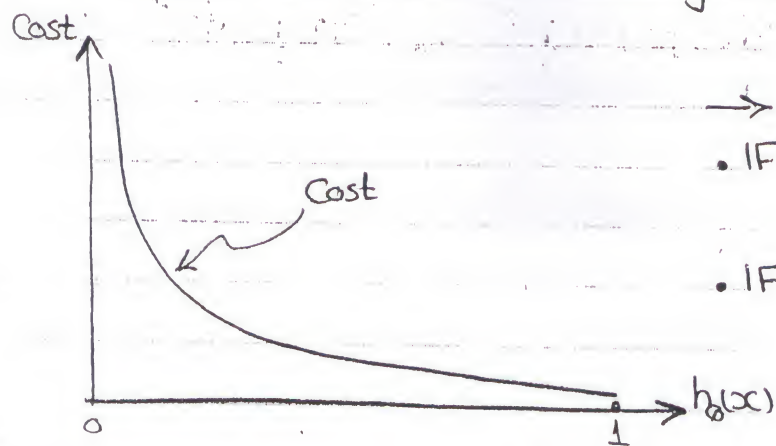
$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\text{Log}(h_{\theta}(x)) & y = 1 \\ -\text{Log}(1 - h_{\theta}(x)) & y = 0 \end{cases}$$

$$y = 1$$

$$y = 0$$

IF $y = 1$

$$\text{Cost}(h_0(x), y) = -\text{Log}(h_0(x))$$



→ Ideal Case $y = h_0(x) = 1$

• IF $y=1$, $h_0(x)=1$

∴ Cost = 0 (No Cost)

• IF $y=1$, $h_0(x) \rightarrow 0$

∴ Cost $\rightarrow \infty$ (Worse Case)

here the point belongs to another class.

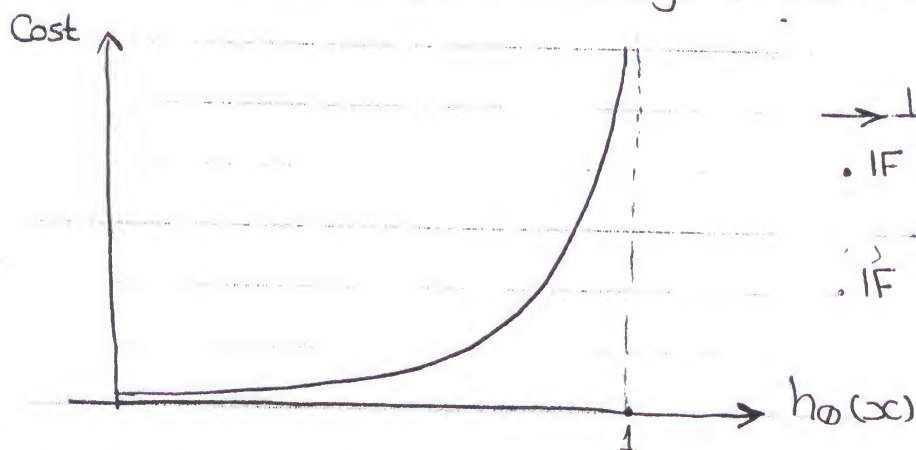
→ Captures intuition that if $h_0(x) = 0$

(predict $P(y=1|x; \theta) = 0$) but $y=1$,

We will penalize Learning algorithm by a very Large Cost.

IF $y = 0$

$$\text{Cost}(h_0(x), y) = -\text{Log}(1 - h_0(x))$$



→ Ideal Case $y = h_0(x) = 0$

• IF $y=0$, $h_0(x)=0$

∴ Cost = 0 (No Cost)

• IF $y=0$, $h_0(x) \rightarrow 1$

∴ Cost $\rightarrow \infty$ (Worse case)

(Large Cost)

Logistic regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\text{Log}(h_{\theta}(x)) & \text{if } y=1 \\ -\text{Log}(1-h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

Simplified Cost Function

$$\rightarrow \text{Cost}(h_{\theta}(x), y) = -y \text{Log}(h_{\theta}(x)) - (1-y) \text{Log}(1-h_{\theta}(x))$$

$$\text{at } y=1 \rightarrow \text{Cost} = -\text{Log}(h_{\theta}(x))$$

$$\text{at } y=0 \rightarrow \text{Cost} = -\text{Log}(1-h_{\theta}(x))$$

$$\rightarrow J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \text{Log}(h_{\theta}(x^{(i)})) + (1-y^{(i)}) \text{Log}(1-h_{\theta}(x^{(i)}))]$$

$$\rightarrow \text{To Fit parameters } \theta : \min J(\theta)$$

$$\rightarrow \text{To make a prediction given new } x:$$

$$\text{Output } h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\rightarrow \text{Gradient Descent}$$

$$\text{Repeat } \{ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \}$$

$$\text{Repeat } \{$$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

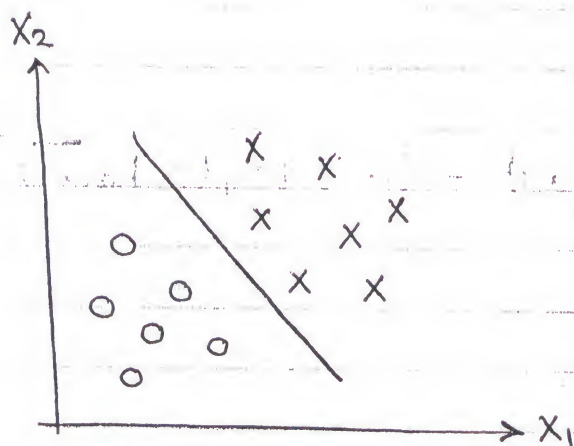
$$\} \quad (\text{simultaneously update all } \theta_j)$$

Multiclass Classification

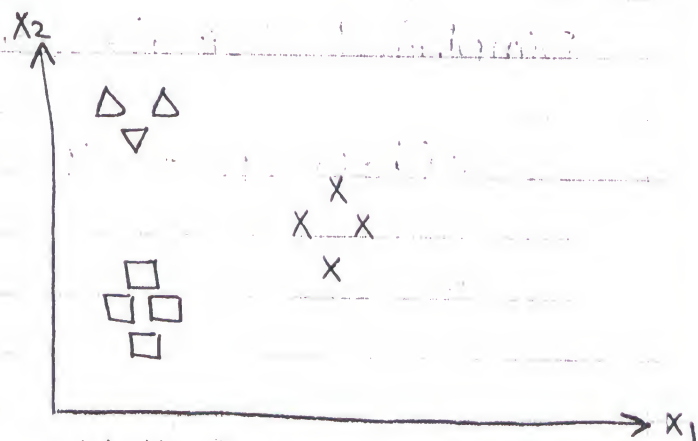
Email Folding / tagging : Work, Frinds, Family, Hobby.

Medical diagrams : Not ill, Cold, Flu.

Weether : Sunny, Cloudy, Rain, Snow



Binary Classification

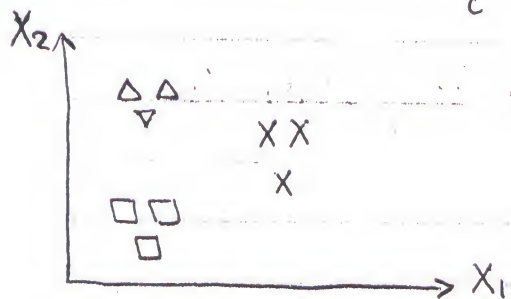


Multi-class Classification

One VS all

- Train a Logistic regression classification $h_{\theta}^{(i)}(x)$ For each class i to predict the probability that $y = i$.

- On a new input x , to make a prediction, pick the class i that maximizes $\max_c h_{\theta}^{(i)}(x)$.



Class 1 : Δ

Class 2 : \square

Class 3 : \times

$$h_{\theta}^{(i)}(x) = P(y=i|x;\theta) \quad (i=1,2,3)$$

